STAT 401 Chapter 3.1–3.4, 3.8, 3.9

Zepu Zhang
September 29, 2010

While we can always fit a simple linear model (SLM) to a $Y \sim X$ data set, and the calculation is easy, this model may not be appropriate for the data set (and the subject behind the data). The "inappropriateness" is usually in the form of violation to one or more assumptions of the model:

1. $$E(Y \mid X) = \beta_0 + \beta_1 X$$
   that is, between $E(Y)$ and X is a linear relationship.
2. $$\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$
   that is,
   (a) ϵ's at different X levels are independent.
   (b) $\text{var}(\epsilon) = \sigma^2$, constant, does not change with X.
   (c) ϵ is normal.

In addition, an important problem to check for is

3. There are important predictors other than X that should have been included in modeling Y.

The most efficient diagnostics are graphics, mainly (1) scatter plot of $Y \sim X$; (2) various plots of the residuals e_i.

Graphical methods are informal, but are really effective and simple. They are usually adequate in discovering problems, although you may need to (or want to) provide something more formal (fancier) to support that a problem that is obvious in the graphics does exist.

"Formal methods" refer to various statistical tests. They are all based on the same idea: find a test statistic whose sampling distribution is known provided a certain assumption (e.g. ϵ is normal) holds, then if the observed value of the statistic is an extreme value in its sampling distribution, it suggests the assumption in question most likely does not hold.

The various problems in a model are intertwined. For example, problem A may obscure problem B, or make an nonexistent problem B appear serious. One fix often alleviates more than one problem.

1 Basic checks on the data

Watch for the following problems in X:

1. Values of X are not distributed in a balanced fashion in the observed range. Use a dot plot.
2. Outlying values of X. Use a dot plot or box plot.

Watch for outlying (x, y) data points. Such points often have both x and y perfectly within their respective ranges but the point (the combination of x and y values) is unusual compared to other data points.

1. $Y \sim X$ scatter plot.
2. $e_i \sim x_i$ plot.

A outlying data point (either x is outlying or the (x, y) combination is outlying) is overly influential on the model estimates. Outlying observations must be treated somehow, but should not be discarded unless there is evidence that the values are erroneous.

2 Is a linear model suitable?

1. Look at a $Y \sim X$ scatter plot. Fig 3.3(a), page 105.
2. Plot e_i against x_i. Symptom: e_i does not appear to fluctuate up and down randomly, but show some pattern of undulation. Fig 3.3(b), page 105. Fig 3.4(b), page 106.

   Caution: an outlier point may cause this effect as well. Fig 3.7, page 109.

Method 1 can be used before model fitting is conducted. By method 2 the problem may be more pronounced. Usually method 1 should be adequate for this problem.

Fixes:
1. Use a nonlinear model. (Not covered.)
2. Transform X. (In a moment.)

Note: when the model is inappropriate, subsequent inferences (Chapter 2) become meaningless. See comment 4, page 127.

3 Diagnostics with residuals

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

The residuals are somewhat like estimates of the random error term (ϵ) in the model, but not quite.

Note that e_i is a random variable if we repeat the data (Y, that is) sampling and model fitting procedure many times.

Properties of the residuals (p. 23–24, 102–103):

1. e_i has a normal distribution with mean 0, because e_i is a linear combination of normal variables Y_i, $\hat{\beta}_0$, $\hat{\beta}_1$, and $E(e_i) = E(Y_i) - E(\hat{Y}_i) = 0$.

2. $\text{var}(e_i)$ is not constant; it varies with X_i. If we work out $\text{var}(e_i)$, we'll see it depends on X_i. (The farther away from the center of X, the larger the $\text{var}(e_i)$.)

3. $\sum e_i = 0$. (Proven earlier.)

4. $\sum X_i e_i = 0$.

5. The e_i's are not independent. This is apparent from the constraints 3 and 4. (They can't be independent, since they have to "cooperate" to satisfy these relations.) If we work out $\text{cov}(e_i, e_j)$, we'll see it's nonzero, which also suggests e_i and e_j are dependent.

6. We call s^2 (or MSE) the "variance" of the residuals. Recall $s^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (e_i - \bar{e})^2}{n-2}$. Remember s^2 is an unbiased estimator of σ^2. However, since the e_i's are not a random sample from a common distribution, it is a little unclear what we mean by the "variance" of the residuals. (They do not have the same variance, according to property 2.)

The properties 1, 2, and 5 will become trivial after we learn multiple regression and use matrix.

When the sample size (n) is reasonably large (and the X values are spread out "nicely"), the residuals are a very effective diagnostic tool—they help to identify violations of the assumptions about ϵ.

Semistudentized residuals (page 103)

$$e_i^* = \frac{e_i - \overline{e}}{\sqrt{s^2}} = \frac{e_i}{s}$$

This is analogous to “standardizing” a normal variable to standard normal. The e_i^* has approximately a t_{n-2} distribution (which is not very different from the standard normal). Because the t distribution is also called “student t” distribution, this operation is called “studentization”. In addition, because the statement $e_i^* \sim t_{n-2}$ is only approximate here, we add “semi-” in front of “studentization”.

3.1 Nonconstancy of error variance

A $e_i \sim x_i$ plot. Fig 3.4(c), page 106. Fig 3.5, page 107.

$Y \sim X$ may also suggest this problem, but less clearly than a $e_i \sim x_i$ plot.

A $e_i \sim \hat{Y}_i$ plot can spot the same problem.

Fix: often calls for a transformation on Y. This operation is called “variance stabilization’.

3.2 Nonindependence of error terms

This is a concern primarily when there is an ordering of the data, for example by time or spatial location.

Plot e_i against the index variable (say time or spatial coordinates). If the residuals show a pattern of undulation, it indicates dependence between the error terms. Fig 3.8, page 109.

Note: such patterns are not visible on a $e_i \sim x_i$ plot (unless X itself is the ordering index variable), because the order of X values is different from the order of time/space.

Fix: Do something about the index variable. See Chapter 12 (not covered in this course).

3.3 Nonnormality of error terms

To check normality of a sample, here $\{e_i\}$, the most commonly used tool is a normal probability plot.

Suppose the residuals are $e_1, e_2, \ldots, e_n$, ordered from small to large. Very roughly speaking, e_i is the $\frac{i}{n}$ quantile of the distribution behind this sample.

However, this is not quite right. For example, is e_n the $\frac{n}{n}$ quantile? No! That would mean e_n is the maximum possible value in the distribution—we have no reason to believe this. So, some tuning (with theoretical justifications) is needed. One tuning is this:

regard e_i as (an estimate of) the

$$\frac{i - .375}{n + .25} \text{ quantile of the distribution of } e$$

(An alternative tuning: regard e_i as the $\frac{i-.5}{n}$ quantile.)

(Recall that the q quantile, written as x_q for example, of a distribution $F(x)$ is such that $F(x_q) = q$, where $F(x)$ is the cdf.)

The normal probability plot is a plot of the $\frac{i-.375}{n+.25}$, $i = 1, \ldots, n$, quantiles of $N(0, 1)$ against $e_1, \ldots, e_n$. That is, it's a plot of corresponding quantiles of the distribution of e and of a standard normal distribution. For this reason, it's called a QQ plot.

If the plot shows (roughly) a straight line, then the distribution of the residuals is normal; if the straight line is $Y = X$, then the distribution of the residuals is standard normal.

Note: It's a general straight line; it does not need to go through $(0, 0)$, and its slope does not need to be 1.

If e is non-normal, the usual pattern is the two ends of the plot deviate from a straight line. From the direction of the deviations, one can tell how the tails of the distribution of e differ from a normal distribution.

Example Say something about the deviation from normal by looking at the normal probability plot. Fig 3.9, page 112.

R functions `qqnorm`, `qqline`, `qqplot`.

Note

1. Normal probability plot is available in possibly all statistical computer software. It's a special case of Q-Q plot. The general Q-Q plot would plot corresponding quantiles of two distributions, or two datasets, or a dataset and a known distribution, in order to compare the distributions.

2. The normality assumption is that ϵ corresponding to any fixed x is normal. Here, e is treated as estimates of ϵ. But e_i correspond to different X values. We actually are assuming independence and constant variance for e, and are pooling the e's at different X and treating them as something like an iid sample of ϵ (whose distribution is the same regardless of the value of X).

   For this reason, independence and constant variance should be checked before normality. Only after the former two checks pass does it make sense to check normality.

3. Why a straight line?

   (1) If two distributions have identical CDF's, then the distributions are the same.

   (2) If corresponding quantiles are all equal, then the two distributions have the same CDF's (hence the distributions are the same).

   (3) If quantiles of distribution 1 are all equal to those of distribution 2 after a linear transform, then the linear transform of distribution 1 is the same as distribution 2.

   (4) A linear transform of a normal distribution is normal.

   Therefore, if the linear transform of e has a standard normal distribution, then e has a (non-standard) normal distribution.

3.4 Omission of important predictor variables

Suppose we've estimated model $Y = \beta_0 + \beta_1 X + \epsilon$. There is another variable, S, that possibly is closely related to Y. To check whether we should include S in the model, plot residuals e_i against s_i. If some pattern appears, then introducing S may "explain" this pattern.

S may well be categorical. For example, S represent two geographical locations. Do the e's at one location differ systematically from those e's at the other location?

Fix: see Chapter 8.

Example Fig 3.10, page 113.

4 Overview of remedial measures

1. Abandon linear regression, develop a more appropriate model.
2. Add predictors ("multiple", rather than "simple", linear regression).
3. Transform X or Y or both.

5 Transformations

The idea is to fit a SLR with transforms of X and/or Y. Interpretation of the model is then in terms of the transformed variables. Hypothesis tests, and confidence intervals are completed on the transformed variables. The purpose of such transformations is to meet the basic assumptions of the model. (Remember, validity of the estimation and inference methods are all contingent upon the validity of the model assumptions.) There are two basic reasons to transform:

1. To linearize a relationship.
2. To achieve normality and constant variance of errors. (Often, fixing one of these problems fixes the other as well.)

case	const	linear	possible fix	examples
	err var?	relation?		
1	yes	no	transform X; don't transform Y	Fig. 3.13, page 130
2	no	no	transform Y	Fig. 3.15, page 132
3	no	yes	transform Y, then (possibly) X	

In case 1, how do we find the "right" transformation?—If the curve looks like that of $f(x)$ (say x^2, $\log(x)$, etc), then try replacing X by $f(X)$.

In case 2, a nonlinear transform of Y "compresses" and "stretches" large and small Y's differently, hence may remedy the uneven spread of Y (i.e. non-constant error variance). Transforming Y will change the curve—a nonlinear relation could become linear, in which case transforming X is not needed.

In case 3, the important thing is to transform Y first so as to stabilize the error variance. This will change the curve, and a linear relation may become nonlinear. Then we're in case 1, so try a transform on X next.

Non-normality of the residuals may be fixed as we fix the non-constant variance problem. We (usually) do not attack non-normality directly.

Drawback of transforming data (X or Y): It becomes harder to interpret your results (in terms of the untransformed variables).

Box-Cox transformations: this is a widely used “family” of transforms. It tries to fix several possible problems at once. However it has its own problems. You should know about this tool and recognize its formula.